Answers to frequently-asked questions for "Incorporating Model Quality Information in Climate Change Detection and Attribution Studies"

Ben Santer, Karl Taylor, Peter Gleckler, Celine Bonfils, Tim Barnett, Dave Pierce, Tom Wigley, Carl Mears, Frank Wentz, Wolfgang Brüggemann, Nathan Gillett, Steve Klein, Susan Solomon, Peter Stott, and Mike Wehner

This paper is embargoed by *Proceedings of the National Academy of Sciences* (PNAS) until 3:00 Eastern Standard Time on Monday, Aug. 10th, 2009.

Q1: Could you briefly describe your previous work with water vapor?

In 2007, we published a paper[1] that looked at the causes of changes in the moisture content of Earth's atmosphere. Since the start of routine satellite-based measurements of atmospheric water vapor in September 1987, atmospheric moisture content has increased markedly. In our 2007 paper, we used rigorous statistical "fingerprint" methods to determine the reasons for the water vapor increase. We showed that human activities were a significant factor in explaining the increase. We demonstrated this by successfully identifying a fingerprint of human activities in the satellite moisture data. The fingerprint was largely due to increases in greenhouse gases caused by the burning of fossil fuels.

Q2: In simple terms, what is climate "fingerprinting"?

"Fingerprinting" involves searching in observed climate records for a pattern of climate change that has been predicted by a computer model (the "fingerprint"). The pattern could be that due solely to natural changes in the Sun's energy output, or to changes in the amount of volcanic dust in the atmosphere. Alternately, it could be a pattern of climate response to human influences, such as human-caused changes in atmospheric levels of greenhouse gas levels. Fingerprint techniques allow researchers to examine a change in some property of the climate system, and then to make rigorous statistical tests of the different possible explanations for that change.

---

[1]Santer, B.D., C. Mears, F.J. Wentz, K.E. Taylor, P.J. Gleckler, T.M.L. Wigley, T.P. Barnett, J.S. Boyle, W. Brüggemann, N.P. Gillett, S.A. Klein, G.A. Meehl, T. Nozawa, D.W. Pierce, P.A. Stott, W.M. Washington, and M.F. Wehner, 2007: Identification of human-induced changes in atmospheric moisture content. *Proceedings of the National Academy of Sciences*, **104**, 15248-15253, doi: 10.1073/pnas.0702872104.

Q3: Did your previous work use results from many different computer models?

Yes. Our 2007 paper used results from 22 different computer models. Model data were taken from the "CMIP-3" archive.[2] This archive contains results from computer model simulations of historical and future changes in climate. These simulations were performed in support of the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC).[3] We relied on climate model output for two different purposes: 1) For estimates of the water vapor "fingerprint" in response to human-caused changes in a variety of different factors[4], and 2) For estimates of purely natural changes in climate.[5] Such estimates of natural variability constitute the "noise" against which we attempt to identify the fingerprint "signal".

Q4: Did your 2007 analysis consider the quality of the 22 models?

Models have different levels of skill in simulating present-day climate. In our 2007 analysis, we did not attempt to judge the relative skill of different models. Instead, we adopted a "one model, one vote" approach. Each of the 22 models contributed equally to the overall estimate of the "model average" fingerprint. Likewise, each model contributed equally to the estimate of the "noise" of natural climate variability.

Q5: How does your new PNAS paper differ from the 2007 analysis?

Some scientists expressed concerns about our "one model, one vote" approach. They argued that we needed to repeat our 2007 analysis with a subset of "better" climate models. This was a reasonable suggestion, so we attempted to identify the "top ten" models out of the full set of 22 we had originally used. We then tried to determine whether we could still identify a human-caused fingerprint if we restricted our attention to these "better" models.

---

[2]CMIP-3 stands for Coupled Model Intercomparison Project, version 3.

[3]IPCC, 2007: Summary for Policymakers. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[4]Examples include changes in greenhouse gases, sulfate and soot aerosol particles, land surface properties, *etc*.

[5]Examples include natural fluctuations in climate associated with such phenomena as El Niños and La Niñas, the Pacific Decadal Oscillation, the North Atlantic Oscillation, and other "modes" of natural climate variability.

Q6: Was it easy to identify the "top ten" models?

No. As other scientists have shown, it is quite difficult to decide which models are the best.[6,7] We considered a total of 70 different objective measures (or "metrics") of climate model performance. Each metric targeted a different aspect of model performance. One group of metrics[8] examined how well the 22 CMIP-3 models captured important features of today's average climate. Another set of metrics[9] focused on the changes in present-day climate that occur over the course of the seasons. The largest group of metrics provided information on model skill in simulating the size and geographical patterns of observed climate variability. We looked at this variability on different timescales – from month to month, year to year, and decade to decade.

We calculated our metrics over a number of regions of climatological interest, and for two different climate variables (water vapor and sea-surface temperature). We found that there was little relationship between model performance in simulating the mean state, the seasonal cycle, and the size and structure of variability. In other words, models that did well in portraying the mean state of water vapor and sea-surface temperature did not necessarily do well in terms of the variability metrics.

Q7: So how did you deal with the difficulty of identifying the "top ten" models?

Any procedure for selecting the "top ten" models involves a large number of subjective decisions. Analysts have to decide which climate variables to look at, which observational datasets to use, which metrics to calculate, which geographical regions and timescales to consider, how to combine information from many different metrics, and ultimately, how to rank the models. Different sets of choices yield different model rankings.

Our strategy was to explore the sensitivity of our fingerprint results to a number of different "model ranking" procedures. We used six different approaches for identifying the "top ten" models. We applied the same approaches to identify six different sets of "bottom ten" models. This enabled us to repeat our fingerprint analysis many times, and to study the sensitivity of our results to climate model quality. We felt that the choice of a single ranking strategy would be very difficult to defend.

---

[6]Gleckler, P.J., K.E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *Journal of Geophysical Research*, **113**, D06104, doi: 10.1029/2007JD008972.

[7]Pierce, D.W., T.P. Barnett, B.D. Santer, and P.J. Gleckler, 2009: Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences*, **106**, 8441-8446.

[8]The "mean state" metrics.

[9]The "annual cycle" metrics.

Q8: And what did you find?

Our bottom-line finding was that "model quality" had very little influence on our ability to identify a human fingerprint in satellite records of water vapor changes. In fact, in every single sensitivity test that we performed,[10] we were able to positively identify a model-predicted fingerprint in the observations. In other words, even if we used the "bottom ten" models to estimate both "climate noise" and a water vapor fingerprint arising from human influences, we were still able to find this fingerprint in the satellite data.

Q9: Why were your fingerprint results so robust?

The water vapor "fingerprint" in response to human-caused changes in greenhouse gases has a very clear and characteristic geographical pattern. This pattern is similar in all models. It shows water vapor increases over the entire global ocean. The largest water vapor increases are over the warmest areas of the tropical oceans. There are smaller increases over the cooler mid- and high-latitude ocean areas. This distinctive pattern is a consequence of very basic physics[11] which is portrayed similarly in all models. This is why we did not see pronounced differences between the water vapor "fingerprint" patterns produced by "top ten" and "bottom ten" models.

Another part of the answer is that the water vapor fingerprint pattern is quite different from the patterns of water vapor changes associated with "climate noise". Again, this is a result that is common to all models. As mentioned above, the fingerprint shows water vapor increases over all ocean areas. The noise patterns do not. In each of the 22 CMIP-3 models, the noise of natural climate variability is characterized by patterns where water vapor increases over some areas and decreases over others. The observed changes in water vapor over the last 21 years are more similar to the fingerprint than to the noise patterns. This similarity between the observations and the "human effects" fingerprint – and the pervasive dissimilarity between the fingerprint and the climate noise patterns – is why our fingerprint results are so robust.

Q10: Do these results suggest that your "model quality" assessment was not useful?

No. Even though our ability to identify a human fingerprint did not depend on whether we used "top ten" or "bottom ten" models, we did find some interesting systematic differences in the fingerprint detection results. When we restricted our attention to the "bottom ten" models, it was consistently somewhat easier to find a human-caused

[10]We performed a total of 144 tests; *i.e.*, our fingerprint detection code was run 144 times.

[11]The Clausius-Clapeyron relationship, which governs the relationship between the temperature at Earth's surface and the moisture-holding capacity of the atmosphere

water vapor fingerprint in the observations. This is because many of the "bottom ten" models underestimate the size of the "real world" climate noise, thus increasing the size of the signal relative to the noise and inflating the statistical significance of the results. This would have been a serious concern if we had been operating close to the fingerprint "detection threshold", where an error in the size of the climate noise could have meant the difference between detecting or not detecting the fingerprint. But for our water vapor problem, we were always <u>well above</u> the "detection threshold" – as our fingerprint results obtained with the "top ten" models clearly showed.

For other climate variables, such as rainfall or ocean heat content, we may be operating much closer to the fingerprint "detection threshold", in which case it may be more important to assess the quality of model-based estimates of climate noise.

<u>Q11: Does your work have any larger implications?</u>

We believe it does.

First, we have shown that the findings of our 2007 paper were robust. We explored the impact of moving from a model democracy ("one model, one vote") to a model meritocracy. Restricting the fingerprint analysis to the "better" models did not affect our ability to identify a human-caused fingerprint in satellite records of water vapor changes. We believe that in the future, such "model quality" assessments are likely to be an integral part of all fingerprint detection work relying on information from large, multi-model archives.

Second, we and many others have clearly shown that model errors are very complex in space and time. Because of this complexity, it's difficult to identify the "top ten" models, even for a very specific application. Our results are relevant to the topical question of whether we should "weight" model projections of future climate changes based on model performance in simulating key features of present-day climate. Most weighting exercises performed to date have used a limited set of "mean climate" metrics to perform the weighting. Our results show that (at least for water vapor and sea-surface temperature) there is little relationship between model skill in simulating today's mean climate and skill in capturing observed variability and the seasonal cycle. This illustrates some of the difficulties that analysts will face in developing objective schemes for weighting model projections of future climate change.[12]

Third, we hope that this study puts to rest the myth that changes in water vapor are purely natural in origin.

Finally, we note that the observed increase in water vapor provides independent

[12]Whether different weighting schemes actually yield different climate change projections is a subject of active research.

evidence of the reality of warming of the lower atmosphere. The observed water vapor increase since 1988 is consistent with pronounced warming of the surface and lower atmosphere, but fundamentally inconsistent with claims (still made by some die-hard skeptics) that the lower atmosphere has cooled over recent decades.

Benjamin D. Santer
Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory
P.O. Box 808, Mail Stop L-103, Livermore, CA 94550, U.S.A.
Tel: (925) 423-3364
Email: santer1@llnl.gov